Decomposition of Large-Scale Semantic Graphs via an Efficient Communities Algorithm

Yiming Yao
(925) 422-1922
yao3@llnl.gov

Semantic graphs have become key components in analyzing complex systems such as the Internet or biological and social networks. These types of graphs generally consist of sparsely connected clusters or "communities" whose nodes are more densely connected to each other than to other nodes in the graph. The identification of these communities is invaluable in facilitating the visualization, understanding, and analysis of large graphs by producing subgraphs of related data whose interrelationships can be readily characterized. Unfortunately, the ability of LLNL to effectively analyze the terabytes of multi-source data at its disposal has remained elusive, since existing decomposition algorithms become computationally prohibitive for graphs of this size. We have addressed this limitation by developing more efficient algorithms for discerning community structure that can effectively process massive graphs.

Project Goals

Current algorithms for detecting community structure are capable of processing only relatively small graphs. The cubic complexity of Girvan and Newman makes it impractical for graphs with more than approximately 10^4 nodes. Our goal for this project was to develop methodologies and corresponding algorithms capable of effectively processing graphs with up to 10^9 nodes. From a practical standpoint, we expect the developed scalable algorithms to help resolve a variety of operational issues associated with the productive use of semantic graphs at LLNL.

Relevance to LLNL Mission

In recent years, LLNL has developed semantic graph technologies capable of fusing disparate facts from diverse sources into massive semantic graphs to facilitate inference of complex and anomalous behaviors embedded within the data. A critical challenge in effectively applying this technology to the Laboratory's mission is to decompose massive graphs into meaningful subgraphs that an analyst can efficiently interrogate to identify these behaviors. This research represents a significant contribution to LLNL's counterterrorism, biodefense, and nonproliferation missions, because efficient decomposition methodologies will provide the foundation for information analysis environments enabling large-scale data mining, and information discovery and visualization.

FY2007 Accomplishments and Results

During FY2007, we completed a graph clustering implementation that leverages a dynamic graph transformation to more efficiently decompose large graphs. In essence, our approach dynamically transforms the graph (or

Table 1. Computation time reduction over Girvan & Newman's method.

Graph	Nodes	Links	T_{gn}	T	R(%)
Erdos972	5488	8972	1180.1	129.6	89.0
Hep-Th	7610	15751	1524.5	670.0	56.1
Kohonen	4470	12720	22.8	8.7	62.0
Power	4941	6594	723.0	550.4	23.9

T_{gn}: Girvan & Newman's time (min); T: our time (min); R = 100(T-T_{gn})/T_{gn}

Table 2. Computation time for parallel graph clustering.

Graph	Nodes	Links	CPUs	T	Q
G10m	10000000	43749984	4	28.9	0.40
G100m	100000000	298437392	32	706.0	0.72
G1000m	1000000000	2624753446	512	710.0	0.78

T: computation time (min); Q: modularity

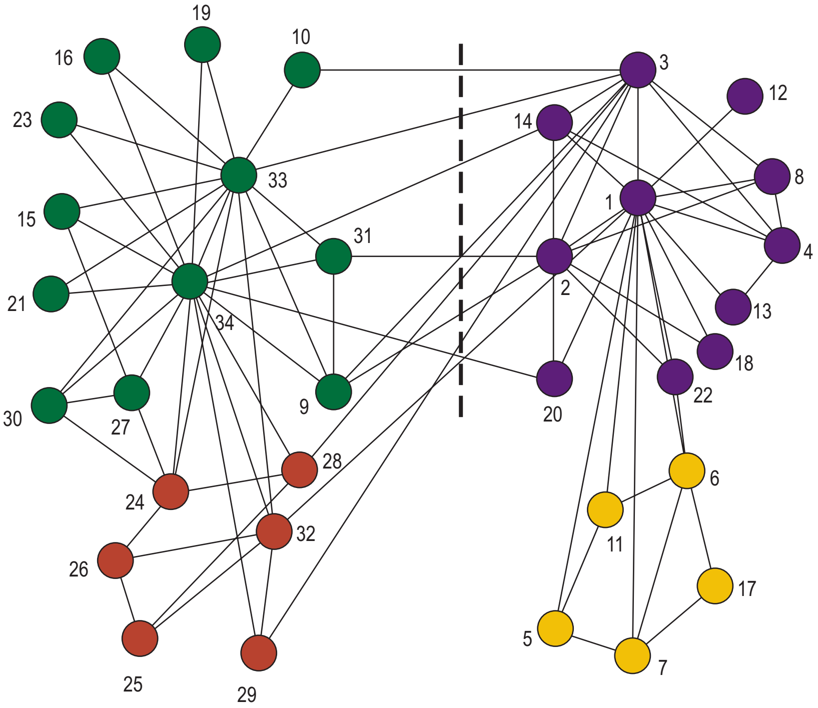
subgraphs) into a tree structure consisting of bi-connected components interconnected by bridge links. This isomorphism allows us to compute edge betweenness, the chief source of inefficiency in Girvan and Newman's decomposition algorithm, much more efficiently, leading to significantly reduced computation time. Test runs on a desktop computer have shown reductions of up to 89% (Table 1).

Our focus this year has been on the implementation of parallel graph clustering on one of LLNL's supercomputers. To achieve efficiency in parallel computing, we have exploited the fact that large semantic graphs tend to be sparse, comprising loosely connected dense node clusters. When implemented on distributed memory computers, our approach performed well on several large graphs with up to one billion nodes, as shown in Table 2. The rightmost column of Table 2 contains the associated Newman's modularity, a metric that is widely used to assess the quality of community structure.

Existing algorithms produce results that merely approximate the optimal solution, *i.e.*, maximum modularity. We have developed a verification tool for decomposition algorithms, based on a novel integer linear programming (ILP) approach, that computes an exact solution. We have used ILP methodology to find the maximum modularity and corresponding optimal community structure for several well-studied graphs in the literature (see figure).

The above approaches assume that modularity is the best measure of quality for community structure. In an effort to enhance this quality metric, we have also generalized Newman's modularity based upon an insightful random walk interpretation that allows us to vary the scope of the metric. Generalized modularity has enabled us to develop new, more flexible versions of our algorithms.

In developing these methodologies, we have made several contributions to both theoretical graph algorithms and software engineering.



Optimal community structure for Zachary's karate club.

Related References

1. Newman, M. E. J., and M. Girvan, "Finding and Evaluating Community Structure in Networks," *Phys. Rev. E* **69**, 026113, 2004.
2. Zachary, W., "An Information Flow Model for Conflict and Fission in Small Groups," *Journal of Anthropological Research*, **33**, pp. 452-473, 1977.
3. Yao, Y., T. L. Hickling, W. G. Hanley, and J. S. Lenderman, "Graph Clustering Evaluation via Integer Linear Programming," *Proceedings of the 2007 International Conference on Data Mining*, pp. 369-375, 2007.